#KiwiIngenuity – Creative Uses of Māori Loanwords in NZE Twitter Posts

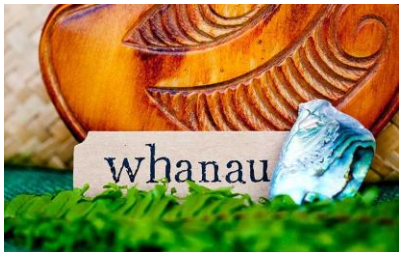
David Trye, Andreea Calude,
Felipe Bravo Marquez & Te Taka Keegan

University of Waikato
University of Chile

1

What can social media tell us about our use of **Māori loanwords** in NZE?

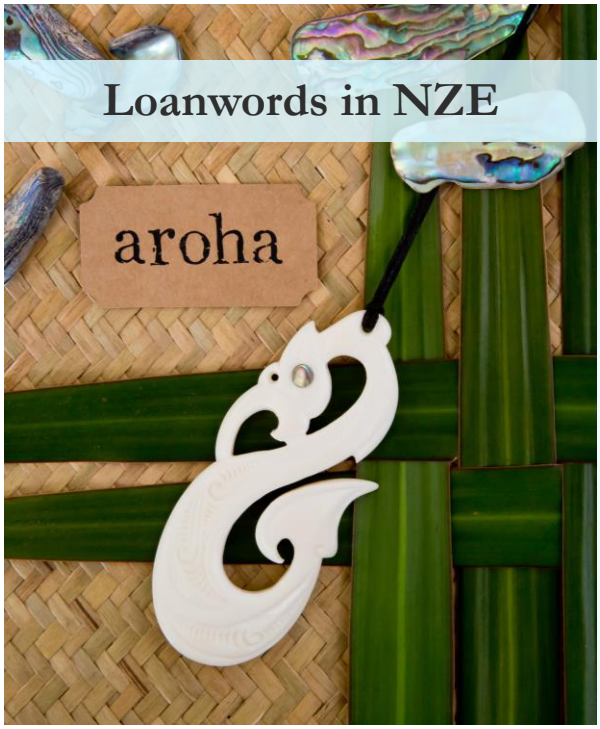
> Kia kaha All Blacks hard. Loss! England well done keeping the the old steam train running on full from the start and the old v stance I'm gud with it being a kiwi Maori #worldinunion #ENGvNZL
>
> 8:27 AM · Oct 27, 2019 · Twitter for iPhone



Background MLT Corpus (Data) Hybrid Hashtags (Findings) Where to next?

2

2



Many genres studied (De Bres, Daly, Macalister, Davies & Maclagan, Degani, Onysko).

For good reason: unusually productive lexical transfer situation from minority language to dominant language.

Two main waves of borrowing (Macalister 2006).

Loanword use remains an increasing trend today, in both types and tokens (especially social culture terms).

Use is skewed across speakers (Māori females lead the change) & across topics (Māori-related topics draw highest counts).

Rationale – Why Twitter?

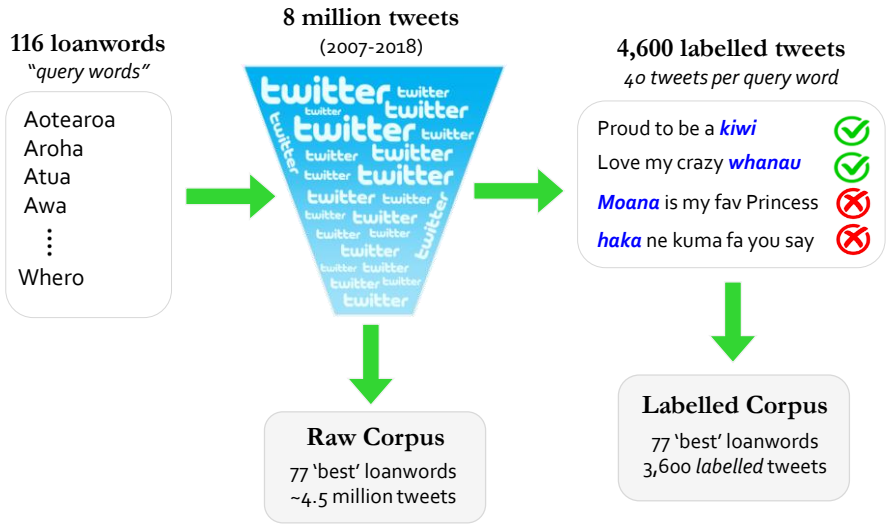
Newspaper Data
- Formal
- Highly Edited
- Prescriptive
- Collaborative
- Normative

Twitter Data
- Formal & Informal
- Not Edited
- Creative
- Single-Authored
- Normative & Non-normative

PLUS
- Cheap to get
- Lots of it

BUT NOISY !!!!

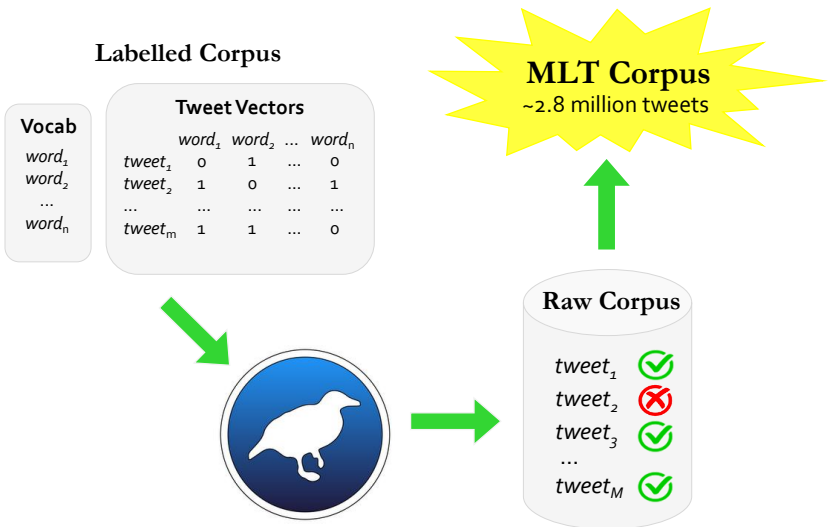
Building the MLT Corpus (1)



Background MLT Corpus (Data) Hybrid Hashtags (Findings) Where to next?

5

Building the MLT Corpus (2)



Background MLT Corpus (Data) Hybrid Hashtags (Findings) Where to next?

6

Machine Learning Input and Output

Training Data (Input)

id	username	timestamp	query word	text	relevance label
757369343642480640	JustStephOK	2016-07-25 12:18	***waiata***	Led the ***waiata*** for the manuhiri at the pōwhiri for new staff for induction week. Was told by the kaumātua I did it with mana & integrity.	**Relevant**

Target Data (Output)

id	username	timestamp	query word	text	prob (rel)
809589244037566460	KUOI_DJ	2016-12-16 15:41	***waiata***	Split Enz—History Never Repeats—***Waiata***	**0.078 (irrelevant)**

Classification Results

	Linear Logistic Regression		
Word *n*-grams	**AUC**	**Kappa**	**F-Score**
1	0.863	0.534	0.801
1, 2	**0.868**	**0.570**	**0.816**
1, 2, 3	0.869	0.560	0.811
1, 2, 3, 4	0.869	0.563	0.813
1, 2, 3, 4, 5	0.869	0.556	0.810

```
@attribute WNGRAM-1-indifferent numeric
@attribute WNGRAM-2-coming-up numeric
...
```

Corpus Overview

	Tokens (words)	Tweets	Tweeters (authors)
Labelled Corpus (Rel.)	~50,000	~2,500	~1,900

PLUS statistically significant, increasing, diachronic trends for 55 of the 77 query words, suggesting that loanword use on Twitter is increasing for most tokens.

9

Hybrid Hashtags

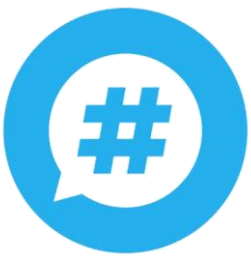
Hashtags containing lexical items from both English and Māori.

#youknowyoureakiwiwhen Christmas Day is spent at the beach 🎄☀️

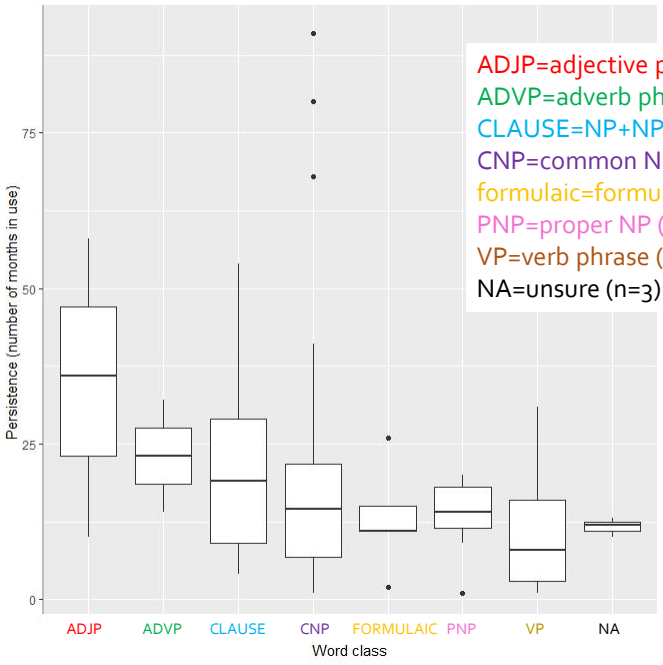
10:40 AM · Nov 22, 2013 · Twitter for iPhone

My son's Korean g'parents, who don't speak English very well, asking us to teach them the reo words for whanau members #LetsShareGoodTeReoStories

8:08 AM · Jan 17, 2018 · Twitter for iPhone



10



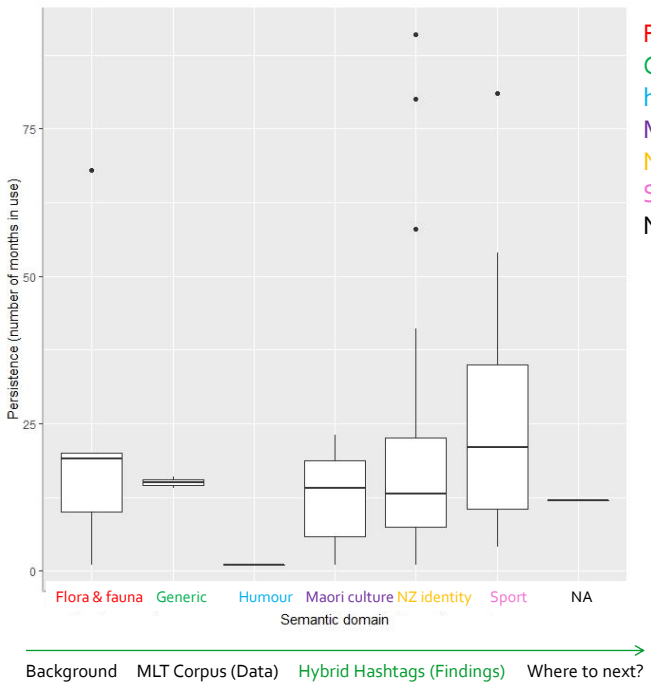
ADJP=adjective phrase (n=3)
ADVP=adverb phrase (n=2)
CLAUSE=NP+NP (n=6)
CNP=common NP (n=42)
formulaic=formula (n=5)
PNP=proper NP (n=7)
VP=verb phrase (n=13)
NA=unsure (n=3)

Syntactic Word Class

#kiwiproud
#proudtobemaori
#proudtobekiwi
#kiwias
#kiwiasbro
#ilovekiwis
#kiwiscanfly
#hakatime
#honorarykiwi
#kiaora4that
#whatkiwisdo
#gothekiwis
#maorifynz
#keepingitreo
#MaoriLanguageWeek
#TreatyofWaitangi

11

11



Flora & fauna (n=5)
Generic (n=3)
humour (n=4)
Maori culture (n=18)
NZ identity (n=35)
Sport (n=15)
NA=unsure (n=1)

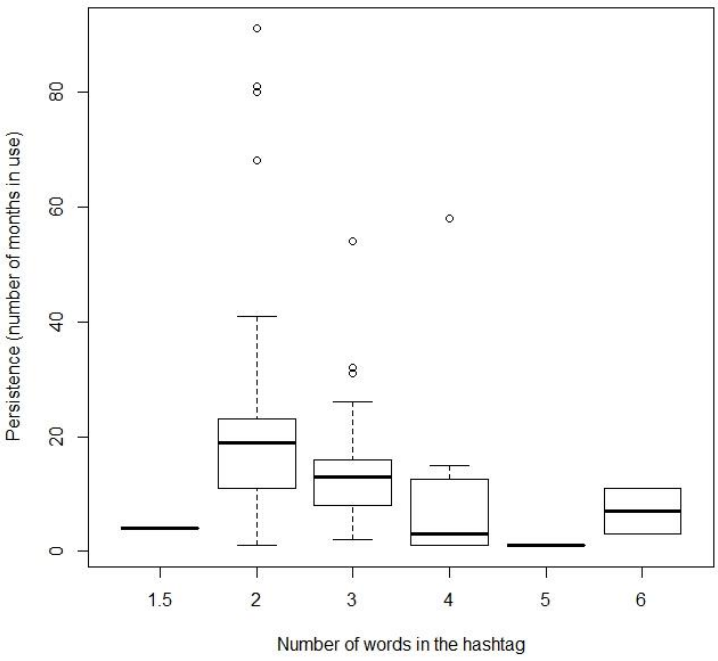
Semantic Domain

#honeyhui
#kiwibird
#kaitime
#kiaora4that
#replacemoviequoteswithkiwi
#NewKiwiBurgerSong
#beingmaori
#goodtereostories
#kiwias
#growingupkiwi
#kiwigold
#lovethehaka

12

Background MLT Corpus (Data) Hybrid Hashtags (Findings) Where to next?

12

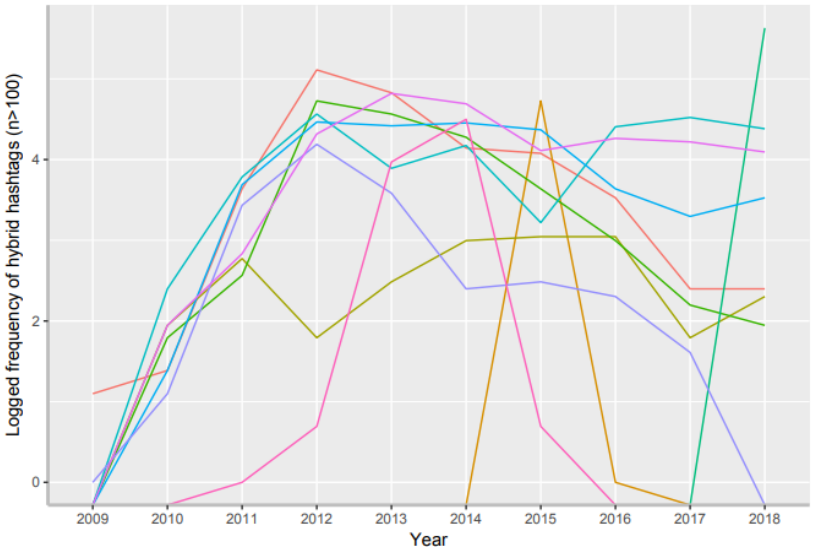


Word Length

13

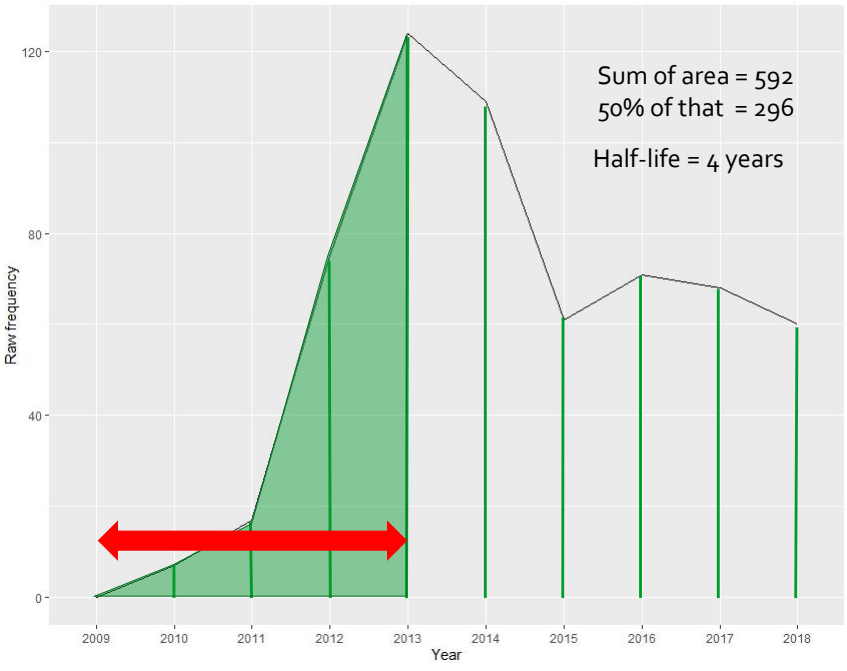
13

Diachronic Trends (Year)



Background MLT Corpus (Data) Hybrid Hashtags (Findings) Where to next?

14



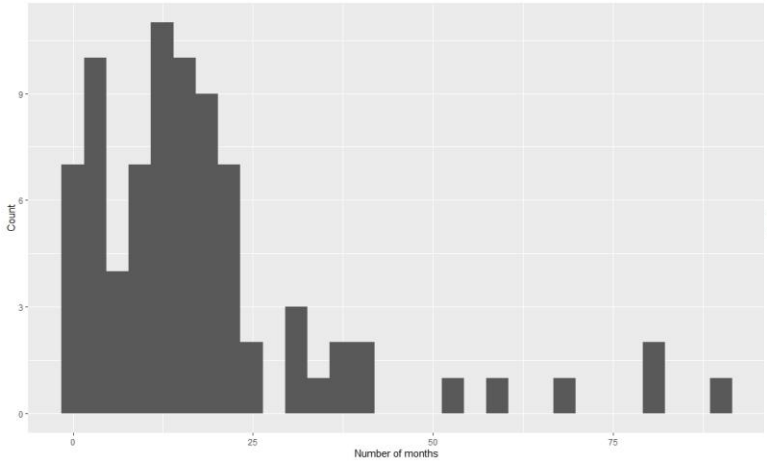
Hashtag Half-life

Background MLT Corpus (Data) Hybrid Hashtags (Findings) Where to next?

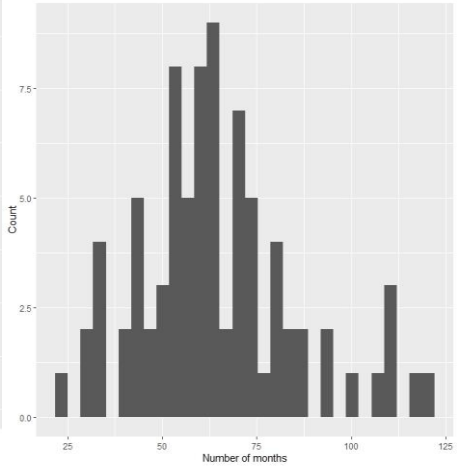
15

15

Comparing Half-life and Persistence



Persistence (num of months it is used)



Half-life (num months it takes to cover half its activity)

Background MLT Corpus (Data) Hybrid Hashtags (Findings) Where to next?

16

```
> model13 = glmer (month_half_life ~ wordclass + semantic_domain + (1|username) + (1|hashtag),
family=poisson, data=sel.me)

     AIC      BIC   logLik deviance df.resid
 22476.9  22563.9 -11224.5  22448.9     3665

Scaled residuals:
     Min       1Q   Median       3Q      Max
-0.64667 -0.00093 -0.00028  0.00012  0.32022

Random effects:
 Groups   Name        Variance Std.Dev.
 username (Intercept) 0.00000  0.0000
 hashtag  (Intercept) 0.07695  0.2774
Number of obs: 3679, groups:  username, 3290; hashtag, 67

Fixed effects:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                   4.1020     0.2927  14.016  < 2e-16 ***
wordclassADVP                 0.4931     0.2837   1.738  0.08218 .
wordclassCLAUSE               0.2164     0.2377   0.910  0.36260
wordclassCNP                  0.3728     0.2056   1.813  0.06984 .
wordclassformulaic            0.4430     0.2624   1.689  0.09131 .
wordclassPNP                  0.5926     0.2483   2.386  0.01702 *
wordclassVP                   0.2015     0.2169   0.929  0.35291
semantic_domaingeneric       -0.2165     0.2913  -0.743  0.45725
semantic_domainhumour        -0.7051     0.2626  -2.685  0.00725 **
semantic_domainMaori_culture -0.2746     0.2270  -1.209  0.22650
semantic_domainNZ_identity   -0.3394     0.2139  -1.586  0.11265
semantic_domainsport         -0.2394     0.2240  -1.069  0.28525
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Predicting Half-life (persistence)

Findings & Future Directions

Twitter is a valuable source of language data for studying loanwords in informal, unedited discourse.

Machine learning techniques and the use of **query words** can be fruitfully used to create mixed-language corpora

One of the most distinctive uses of loanwords in these data appears to be within **hybrid hashtags**.

These hybrid hashtags are **linguistically diverse**:
- They are different lengths (2-6 words)
- Their word class varies (nouns, adj, adv, verb, clauses)
- They pertain to different semantic domains (including NZ identity, Māori culture, sport & humour).

Hybrid hashtags are comprised of many English word-types but **only a handful of loanword-types**
From the 77 query words searched → 81 hybrid hashtags → 9 loans
Partaking in hybrid hashtags may be another predictor of **entrenchment**.

- **Position** of the hybrid hashtag within the tweet
- Number of **followers** for each user

Thank you

- ✉ davidtrye@gmail.com
- ✉ andreea@waikato.ac.nz
- 🔗 https://kiwiwords.cms.waikato.ac.nz/

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

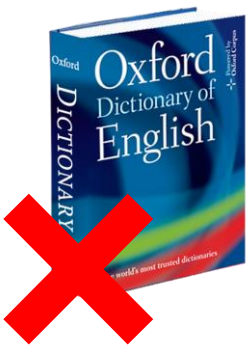
the ROYAL SOCIETY of NEW ZEALAND
TE APĀRANGI

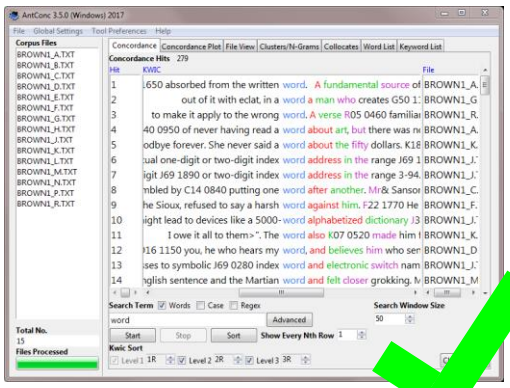
MARSDEN FUND
TE PŪTEA RANGAHAU A MARSDEN

19

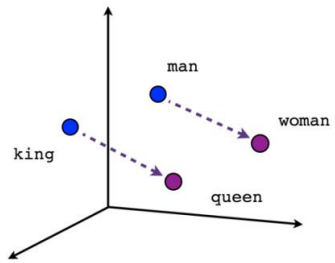
Word Embeddings

Distributional Hypothesis:
"You shall know a word by the company it keeps"





Word2Vec

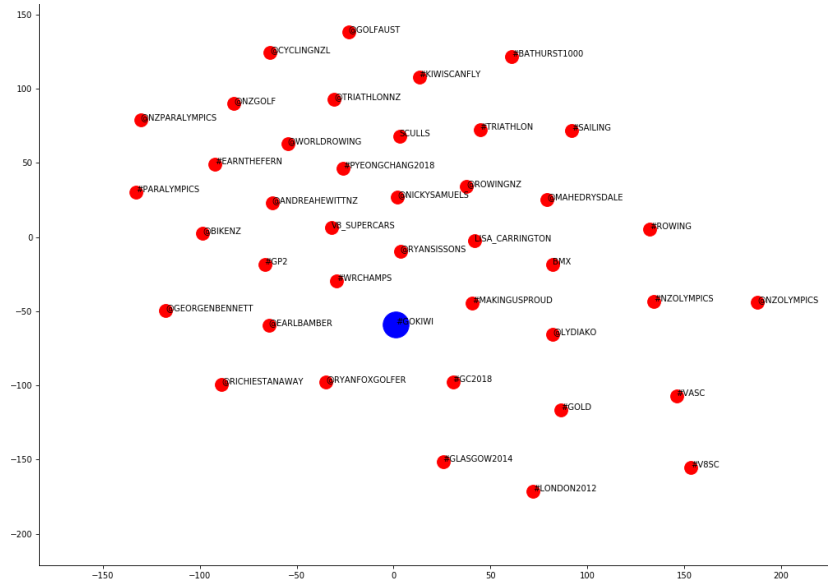


Similar words receive similar vectors, and are situated closer together in *n*-dimensional space

20

Word Embedding Plot for #GoKiwi

Sport

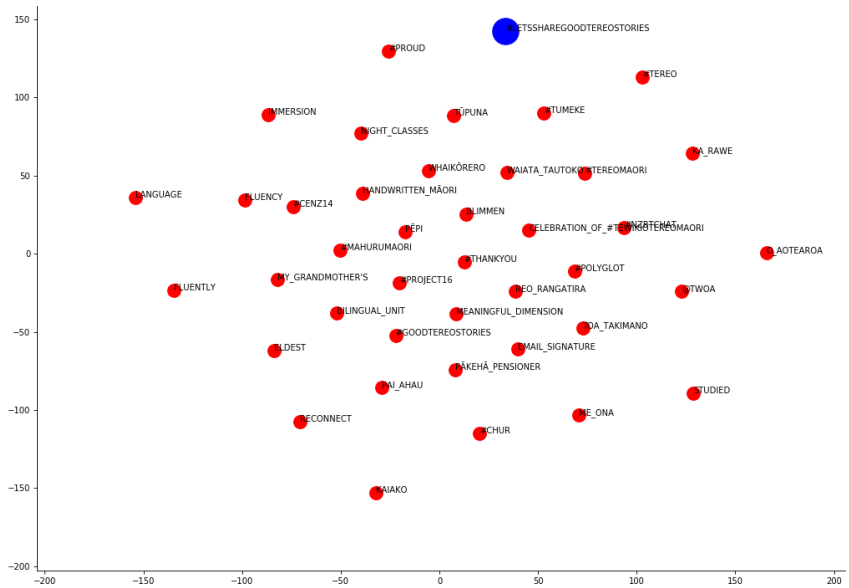


21

21

Word Embedding Plot for #LetsShareGoodTeReoStories

Māori Culture



22

22

Recent Corpora

Matariki Corpus
(with Sally Harper, Steven Miller & Hēmi Whaanga)

- » 2007-2016
- » ~ 91, 958 words
- » ~ 194 articles
- → 2,673 loanword tokens /
- → 282 loanword types, rate 29/1,000 words

National Science Challenge Corpus
(with Louise Stevenson, Hēmi Whaanga & Te Taka Keegan)

- » Snapshot in Jan 2018
- » ~ 1.5 million words
- » ~ 12 websites & 11 Twitter feeds

Māori Language Week Corpus
(with Katie Levendis)

- » 2008-2017
- » ~ 108,925 words
- » ~ 290 articles
- → 3,795 loanword tokens /
- → 186 loanword types, rate 35/1,000 words

Māori Loanword Twitter (MLT) Corpus
with David Trye, Felipe Bravo Marquez & Te Taka Keegan

Background MLT Corpus (Data) Hybrid Hashtags (Findings) Where to next?

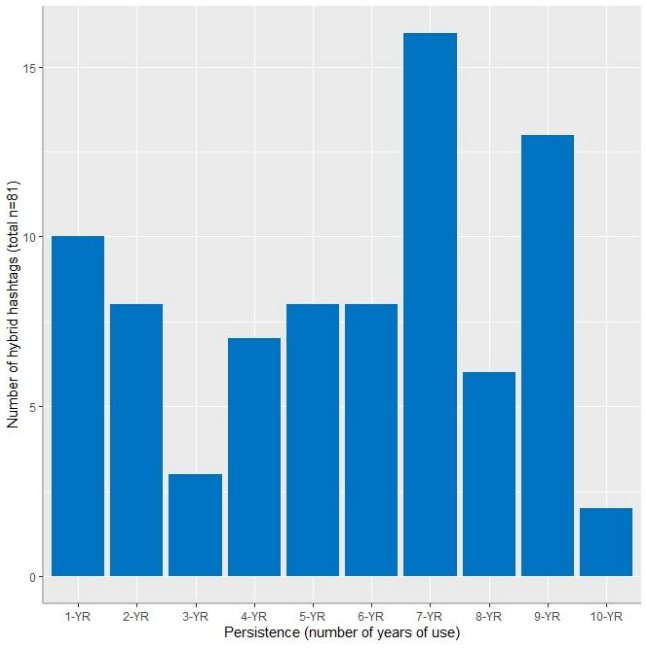
23

Comparisons of Loanword Use with Other Corpora (Top 10)

Wellington Written Corpus (1993)	Matariki Corpus (2007-2016)	Māori Lang. Week Corpus (2008-2017)	Māori Loanword Twitter (2007-2018)	Hybrid Hashtags (2009-2018)
Maori	Māori	Māori	Kiwi(s)	Kiwi(s)
Pakeha	Matariki	te reo	Māori	Māori
Kiwi	marae	iwi	haka	haka
marae	puanga	reo	kia ora	(te) reo
kakapo	kapa haka	whanau	Whānau	hui
tangata	te reo	marae	Aotearoa	Waitangi
tiriti	whānau	kapa haka	kia kaha	Aotearoa
pa	waka	Pakeha	mōrena	kai
whenua	hangi	Kiwi	te reo	kia ora
aroha	iwi	kia ora	whare	

Background MLT Corpus (Data) Hybrid Hashtags (Findings) Where to next?

24



#kiwasbro
#kiwiberries
#kiwibird
#kiwigirl
#kiwigold
#kiwilegend
#kiwilife
#kiwiproblems
#kiwisrock
#kiwistyle
#gothekiwis
#kiwiaccent
#kiwias
#kiwifruit
#kiwimusic
#kiwipride
#kiwiscanfly
#proudtobeakiwi
#maoriculture
#NZMaori
#thehaka
#TreatyofWaitangi
#UptheKiwis
#MaoriLanguageWeek
#WaitangiDay

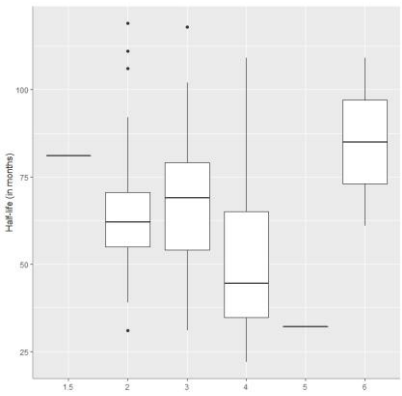
Which Hashtags "Survive" Longer?

25

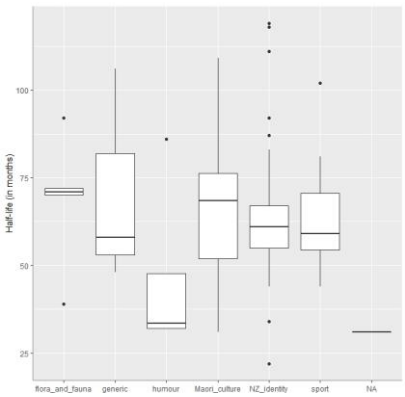
Background MLT Corpus (Data) Hybrid Hashtags (Findings) Where to next?

25

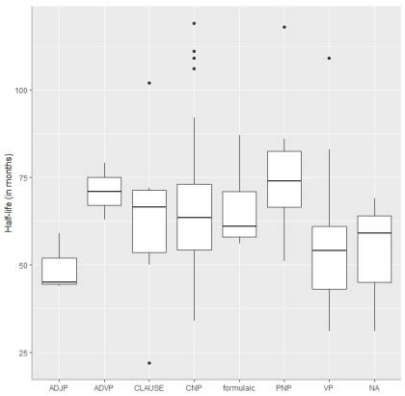
Using Half-lifes



Length of hashtag (number of words)



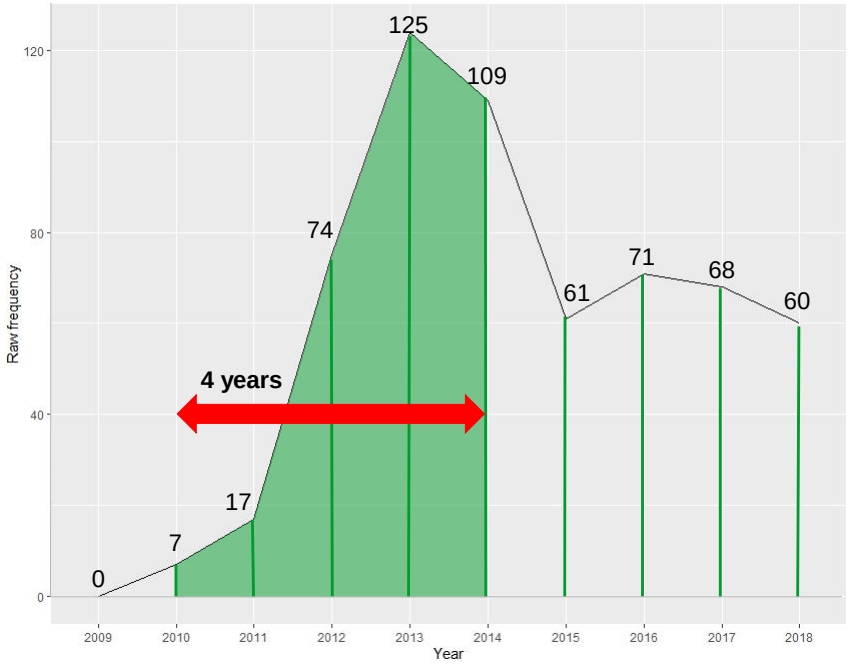
Semantic Domain



Word Class

Background MLT Corpus (Data) Hybrid Hashtags (Findings) Where to next?

26



Half-life = 4 years

Sum of area = 0+7+17+74+125+109+61+71+68+60=592

50% of area = 592/2 = 296

0+7+17+74+125 =

Half-life = 4 years

27